Data Analytics

CareerFoundry
Data Immersion
Achievement 1: Final Exercise

Preparing for Influenza Season

Interim Report

Jessica Duong
January 2026

TABLE OF CONTENTS

PROJECT OVERVIEW

Goal

To help a medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis. The analysis will help plan for influenza season, a time when additional staff are in high demand. The final results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

Motivation

The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.

Objective

Estimate which states and time periods will require additional medical staffing during influenza season using mortality, population, and vaccination data.

Scope

The agency covers all hospitals in each of the 50 states of the United States plus the district of Columbia, and the project will plan for the upcoming influenza season.

Stakeholder Identification

- Medical agency frontline staff (nurses, physician assistants, and doctors)
- Hospitals and clinics using the staffing agency's services
- Influenza patients
- Staffing agency administrators

HYPOTHESIS

Research Question

Do states with higher percentages of elderly population (ages 65+) experience higher influenza death rates, and does vaccination coverage among seniors mitigate this risk?

OVERVIEW OF THE DATA

Data Sets Used

1. Influenza Deaths by Geography

Summary: CDC state-level influenza deaths by age group, 2009 to 2017.

Limitations: Only primary causes of death are reported; comorbidities are unavailable. Small counts (<10) were suppressed, and Puerto Rico was excluded due to missing data.

Source: https://wonder.cdc.gov/ucd-icd10.html

2. Population Data by Geography, Time, Age and Gender

Summary: State-level population counts by age and gender, 2009 to 2017 (U.S. Census).

Limitations: Population counts are survey estimates and may include coverage gaps.

Source: This data is collected through surveys by the United States Census Bureau.

3. Vaccination Coverage Data (65+ Population)

Summary: State-level influenza vaccination coverage for ages 65+, 2009 to 2016 (CDC).

Limitations: Population counts are survey estimates and may contain coverage gaps due to infrequent collection.

Source: https://data.cdc.gov/Flu-Vaccinations/Influenza-Vaccination-Coverage-for-All-Ages-6-Mont/vh55-3he6/about_data

4. CDC Influenza Visits Dataset (Healthcare Provider Capacity)

Summary: State-level counts of healthcare providers reporting influenza-like illness (ILI).

Limitations: Represents only reporting facilities.

The Integrated Data Set

- The influenza deaths, population data, and vaccination data sets were combined using the concatenated variable **State-Year** as a key identifier
- The age groups were combined into 2 groups:
  - High Risk (65+ Years of Age)
  - Low Risk (Under 65 Years)
- Influenza deaths were calculated as death rates per 100,000 population to normalize across different population sizes
- Vaccination rates were included to test intervention effectiveness

- Final dataset: 458 state-year observations across 50 states + District of Columbia from 2009-2017

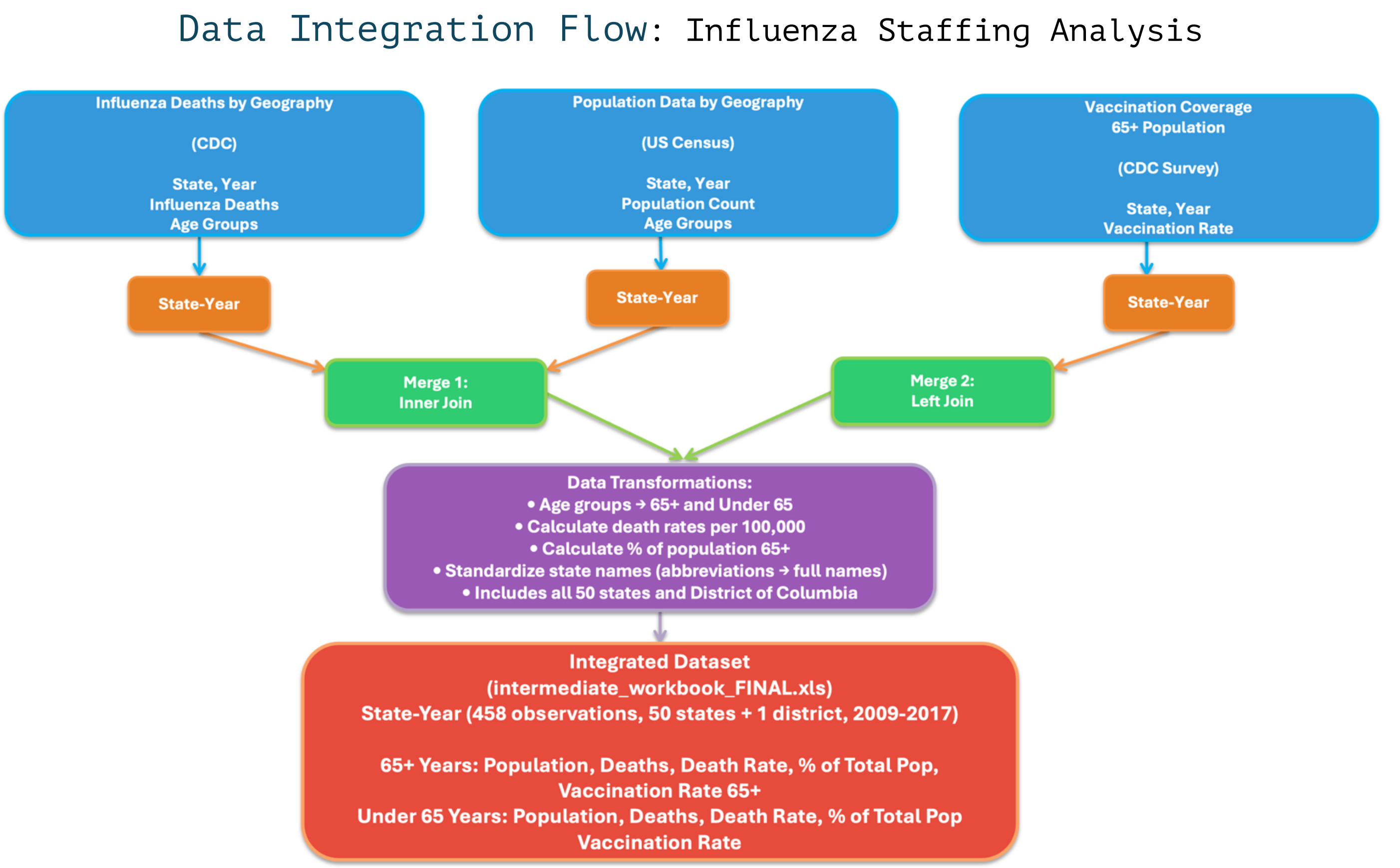


Figure 1: Key: State-Year = Concatenated identifier (e.g., 'Alabama_2009') used to merge datasets.

DESCRIPTIVE ANALYSIS

Summary Statistics

Variable	65+ Years	Under 65 Years	65+ Years	Under 65 Years	65+ Years	Under 65 Years
Influenza Deaths			**Population Count**		**Death Rate per 100,000**	
Mean	891.69	491.91	892,863	5,678,013	120.78	23.07
Median	571.25	438.50	639,195	4,028,026	114.47	12.48
Max	5,694	1,311.50	5,115,069	33,840,957	304.19	100.28
Min	162	432	62,486	457,102	52.02	2.58
Q1	256.12	432	287,339	1,783,456	96.15	7.63
Q3	1,115.38	484.12	1,014,853	6,740,277	132.79	30.14
IQR	859.25	52.12	727,513	4,956,821	36.64	23.03
Standard Deviation	976.17	123.58	901,826	6,057,501	36.64	23.03

Table 1: Summary of descriptive analytical measures for the integrated data, grouped by high risk (65+ Years) and low risk (Under 65) comparing population counts and influenza deaths.

Distribution of Influenza Deaths: Key Observations

- 65+ death rates are ~5x higher than under 65
- Deaths are concentrated in larger, older states
- Variability is substantially higher in the 65+ group

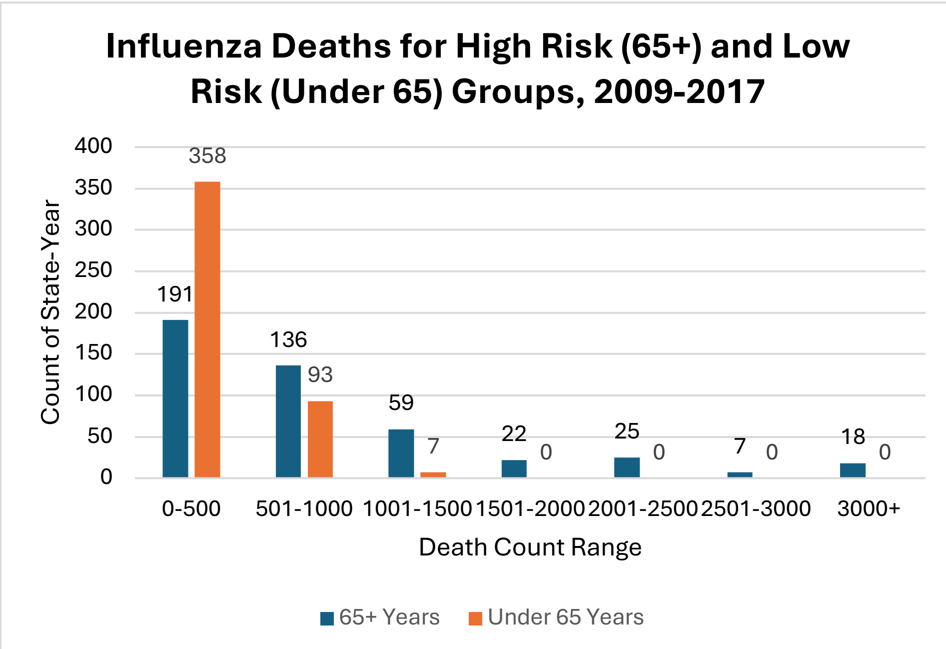


Figure 2: *Distribution of influenza deaths by age group, showing substantially higher counts among seniors.*

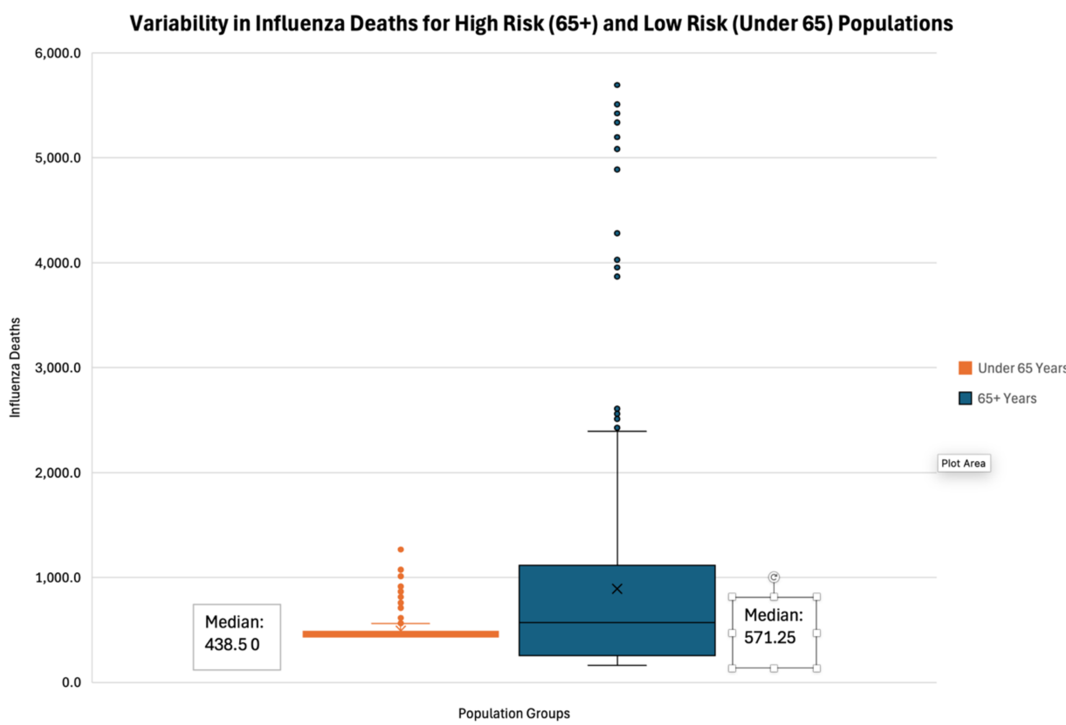


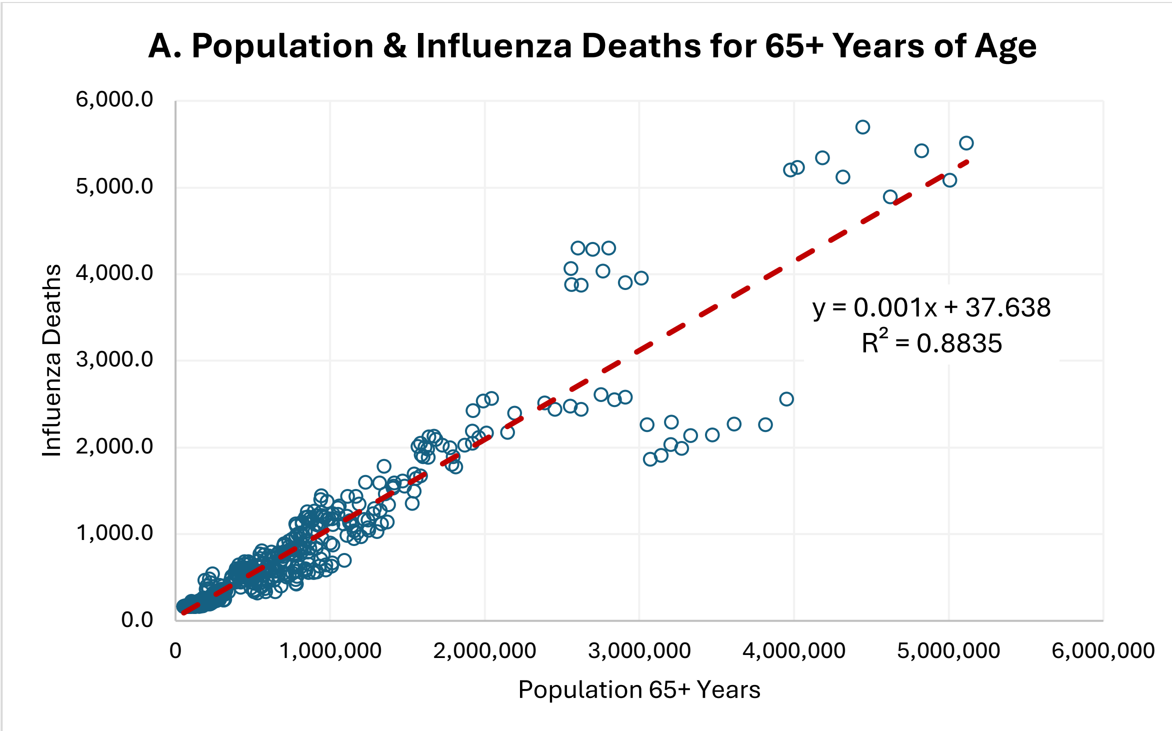
Figure 3: *Box plot comparing variability in the influenza deaths for high risk (65+ years) and low risk (Under 65) populations, 2009-2017.*

Variability in Deaths

Deaths vary much more in the 65+ population (IQR 859 vs 52), indicating that state size and elderly population drive most variability.

Correlation Between Population and Deaths

Influenza deaths scale strongly with population size, especially among seniors ($R^2 = 0.88$).



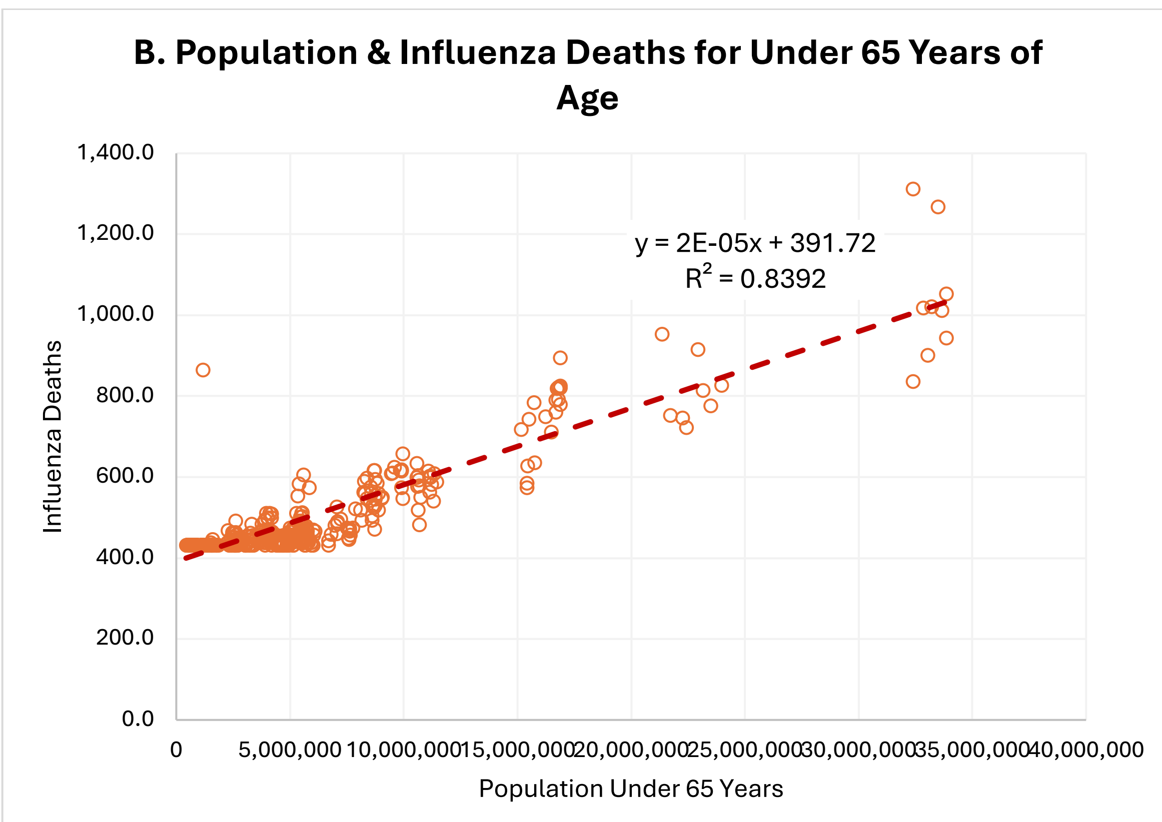


***Figure 4**: Deaths scale with population, strongest among seniors.*

Vaccination Coverage vs Mortality

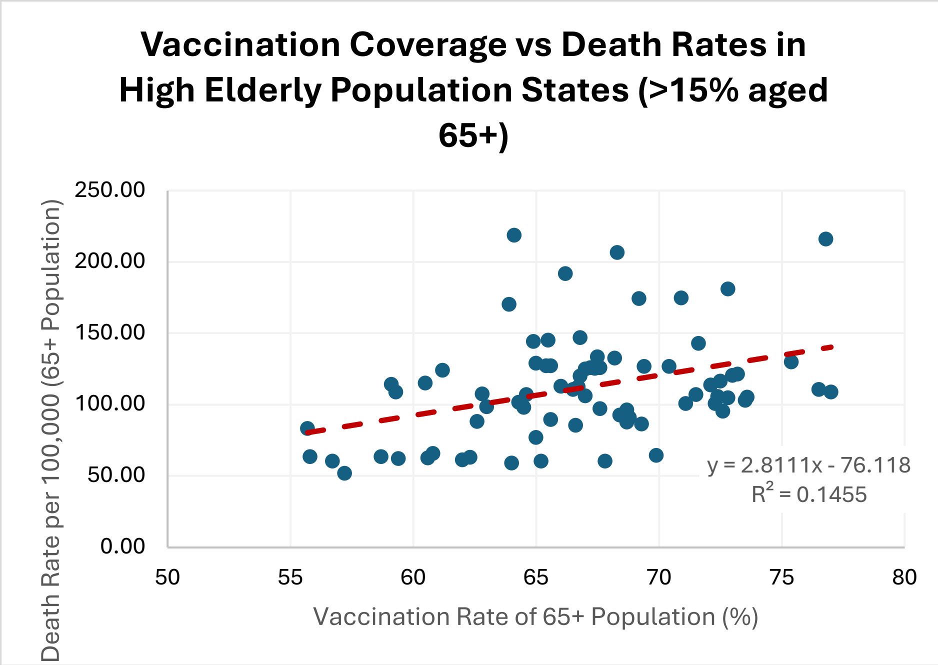


Figure 5: *Weak relationship between vaccination coverage and mortality ($R^2 = 0.15$).*

Vaccination Coverage	Number of Observations	Avg Death Rate (per 100k)	Median Death Rate (per 100k)
Low (<65%)	150	125.22	111.86
Medium (65-70%)	142	114.61	115.73
High (>70%)	113	127.23	122.16

Table 3: *Death rates by vaccination coverage level in high elderly population states (>15% aged 65+), 2009-2017.*

Vaccination Coverage Analysis

Vaccination rates show minimal impact on state-level death rates and are not reliable predictors of staffing demand.

Seasonal Patterns & Provider Capacity

- Peak Season: December to March (highest demand)
- Moderate Season: October to November
- Current providers distributed based on historical reporting patterns. Reallocation model suggests redistributing to better match vulnerable population concentrations.

Staffing Implications:

- States with larger elderly populations require proportionally more medical resources, with needs following reliable linear patterns
- Vaccination rates should not be used to reduce staffing allocation in any state.
- Medical staffing needs fluctuate 5-7x between peak (Dec-March) and low seasons.

INFERENTIAL ANALYSIS

Statistical Hypothesis Testing

Null Hypothesis (H_0): High-risk (65+) and low-risk (under 65) populations have the same influenza death rates per 100,000 population.

Alternative Hypothesis (H_1): High-risk populations (65+) have higher influenza death rates per 100,000 population than low-risk populations.

Statistics	65+ Years	Under 65 Years
Mean Death Rate	120.78 per 100,000	23.07 per 100,000

Test Result	Value
P-value (one-tail)	< 0.0001
t Critical (one-tail, α=0.05)	1.647

NEXT STEPS

Immediate Priorities

- Segment states by elderly population share to prioritize staffing
- Analyze seasonal trends to identify peak months by state
- Incorporate visit/lab data to estimate patient volume (not just deaths)

Enhanced Analysis Considerations

1. **Staffing Requirement Modeling**
   - Build predictive models to forecast staffing demand using population size and historical mortality
   - Define staffing thresholds based on surge patterns
2. **Visualization Development**
   - Develop Tableau dashboards for risk maps, seasonality, and staffing recommendations

Data Quality Improvements

- Document excluded states
- Address missing vaccination data
- Validate results with additional sources

APPENDIX

Data Quality Documentation

Coverage of Final Dataset

- 50 states + District of Columbia with reliable, reportable data
- 458 state-year observations (2009-2017)
- 88.4% vaccination data completeness (11.6% missing, primarily 2017)
- Vaccination dataset: 423 observations after removal of missing records.
- Represents approximately 98% of U.S. population

Alternative Age Group Analysis

Exploratory analysis showed that when we stratify populations into age groups as older or younger than 55 years of age, the correlation of influenza deaths remains strong for 55+ years of age ($R^2 = 0.907$) and moderates slightly for those under 55 years of age ($R^2 = 0.703$).

This suggests that the vulnerable population may extend beyond just those 65+, and individuals in the 55-64 age range also show elevated risk. Future analysis could explore this expanded definition of "high-risk" population to improve staffing models.

Technical Notes

Data Processing:

- State name standardization: Converted all state abbreviations (AL, AZ, etc.) to full names for consistent merging
- Rate calculations: All death rates calculated as (Deaths / Population) × 100,000 for standardization
- During initial raw data cleaning, suppressed values were conservatively imputed at 4.5 deaths to enable exploratory analysis
- However, the final integrated dataset used for analysis (Figures 2-4, Tables 1-2) includes only state-year observations with reportable death counts (≥10 deaths per CDC privacy thresholds)
- Data transformation during the merge process resulted in removal of imputed low-count observations, ensuring statistical reliability

Statistical Methods:

- Welch's t-test used instead of Student's t-test due to unequal variances between groups
- One-tailed test appropriate given directional hypothesis (65+ expected to be higher)
- Alpha level of 0.05 provides 95% confidence threshold